PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-229026
(43)Date of publication of application : 24.08.2001

(51)Int.Cl. G06F 9/44
G06N 1/00

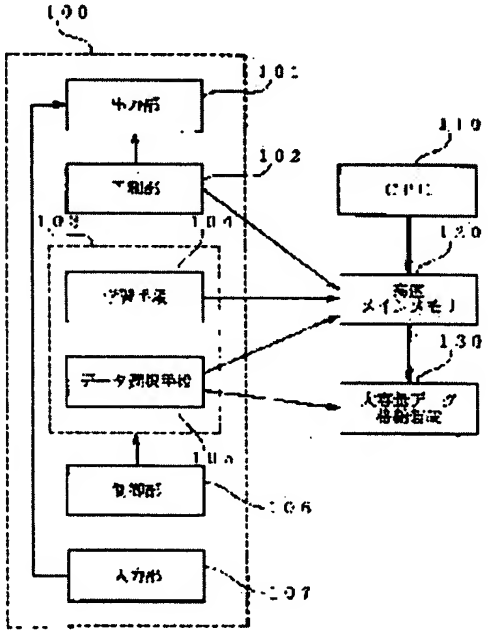
(21)Application number : 2000-104760
(22)Date of filing : 06.04.2000

(71)Applicant : NEC CORP
(72)Inventor : MAMIZUKA HIROSHI
ABE NAOKI

(30)Priority
Priority number : 11350834 Priority date : 09.12.1999 Priority country : JP

(54) KNOWLEDGE DISCOVERING SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a general knowledge discovering system for efficiently realizing knowledge discovery by selectively sampling only data whose information quantity is large from a data base.
SOLUTION: A learning means 104 allows a low rank learning algorithm inputted through an input part 107 to learn plural partial samples prepared by sampling from data stored in a high speed main memory 120, and obtains plural hypotheses. A data selecting means 105 estimates the information quality of each candidate data read from a large capacity data storing device 130 by using the hypotheses, and additionally stores only the data whose information quantity is large in the high speed main memory 120. A control part 106 repeats the processing prescribed number of times, and stores the obtained final hypothesis. Then, a predicting part 102 predicts a label value by the final hypothesis for data inputted to the input part 107 whose label is unknown, and an output part 101 outputs this predicted value.



LEGAL STATUS

[Date of request for examination]	06.04.2000
[Date of sending the examiner's decision of rejection]	30.07.2002
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	
[Date of registration]	
[Number of appeal against examiner's decision of rejection]	2002-16528
[Date of requesting appeal against examiner's decision of rejection]	29.08.2002
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－229026
（P2001－229026A）
(43)公開日　平成13年８月24日(2001.8.24)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　9/44	５８０	Ｇ０６Ｆ　9/44	５８０Ａ	
Ｇ０６Ｎ　1/00		Ｇ０６Ｎ　1/00		

審査請求　有　　請求項の数15　ＯＬ　（全 10 頁）

(21)出願番号　特願2000－104760(P2000－104760)

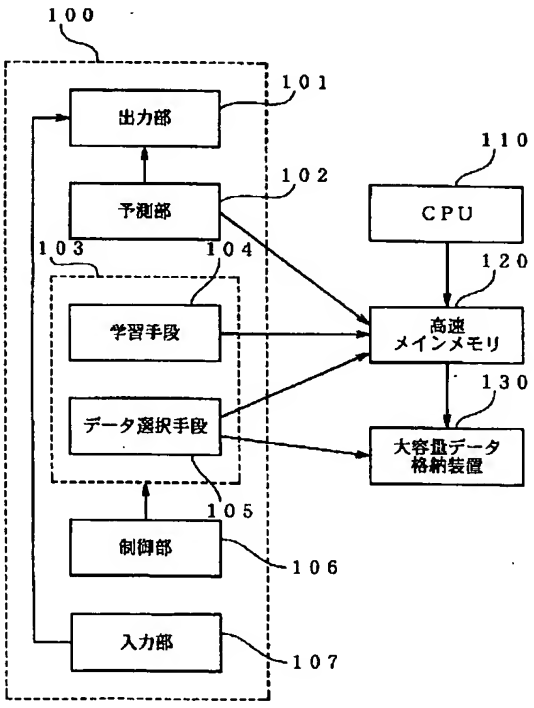
(22)出願日　平成12年４月６日(2000.4.6)

(31)優先権主張番号　特願平11－350834
(32)優先日　平成11年12月９日(1999.12.9)
(33)優先権主張国　日本（ＪＰ）

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　馬見塚　拓
東京都港区芝五丁目７番１号　日本電気株式会社内
(72)発明者　安倍　直樹
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100108578
弁理士　高橋　詔男　（外３名）

(54)【発明の名称】　知識発見方式

(57)【要約】
【課題】　データベースから情報量の大きいデータのみ選択的にサンプリングし、効率的に知識発見を行なう汎用的な知識発見方式を提供する。
【解決手段】　学習手段１０４は、高速メインメモリ１２０に格納されているデータからサンプリングによって作成した複数の部分サンプルを、入力部１０７を介して入力した下位学習アルゴリズムに学習させ、複数の仮説を得る。データ選択手段１０５は、この仮説を用いて、大容量データ格納装置１３０より読み出された候補データ各々の情報量を推定し、情報量の大きいデータのみを高速メインメモリ１２０に追加格納する。制御部１０６は、上記の処理を所定の回数繰りし、得られた最終仮説を格納する。そして、予測部１０２は入力部１０７へ入力されたラベル未知のデータに対し最終仮説によりラベル値を予測し、出力部１０１はこの予測値を出力する。



1

【特許請求の範囲】

【請求項１】　大容量データ格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、

学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、

前記メインメモリに格納されているデータからサンプリングして作成した複数の部分データ集合を前記学習アルゴリズムへ訓練データとして入力して学習させ、複数の仮説を得る学習手段と、

該学習手段により得られた複数の仮説を用いて前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を１つあるいは複数選択し、前記メインメモリに追加して格納するデータ選択手段と、

前記学習手段と前記データ選択手段による選択的なデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリに格納させる制御手段と、

前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段と、

を有することを特徴とする知識発見方式。

【請求項２】　前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに追加して格納することを特徴とする請求項１に記載の知識発見方式。

【請求項３】　前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに追加して格納し、

前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする請求項１に記載の知識発見方式。

【請求項４】　前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに追加して格納することを特徴とする請求項１に記載の知識発見方式。

2

【請求項５】　前記データ選択手段は、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、前記予測値の分散の小さい候補点を１つまたは複数選択して前記メインメモリの格納データから削除することを特徴とする請求項１または２に記載の知識発見方式。

【請求項６】　前記データ選択手段は、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、多値予測の場合、前記予測値のマージンの大きい候補点を１つまたは複数選択し、実数値予測の場合、前記予測値の重み付き分散値の小さい候補点を１つまたは複数選択し、前記メインメモリの格納データから削除することを特徴とする請求項１または３に記載の知識発見方式。

【請求項７】　前記データ選択手段は、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、多値予測の場合、前記予測値の誤差マージンの大きい候補点を１つまたは複数選択し、実数値予測の場合、前記予測値の重み付き予測誤差の小さい候補点を１つまたは複数選択し、前記メインメモリの格納データから削除することを特徴とする請求項１または４に記載の知識発見方式。

【請求項８】　大容量データ格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、

学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、

前記メインメモリに格納されたデータを訓練データとして入力して学習し、学習した仮説を大容量データ格納装置もしくはメインメモリに格納する学習手段と、

該学習手段により得られた過去の複数の仮説を用いて、前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を複数選択し、前記メインメモリに格納するデータ選択手段と、

前記学習手段と前記データ選択手段におけるデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリもしくは大容量データ格納装置に格納させる制御手段と、

前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段と、
を有することを特徴とする知識発見方式。

【請求項９】　前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに格納することを特徴とする請求項８に記載の知識発見方式。

【請求項１０】　前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに格納し、
前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする請求項８に記載の知識発見方式。

【請求項１１】　前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに格納することを特徴とする請求項８に記載の知識発見方式。

【請求項１２】　大容量データ格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、
学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、
前記メインメモリに格納されているデータからサンプリングして作成した複数の部分データ集合を前記学習アルゴリズムへ訓練データとして入力して学習させ、複数の仮説を得る学習手段と、
該学習手段により得られた過去の複数の仮説を用いて、前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を複数選択し、前記メインメモリに格納するデータ選択手段と、
前記学習手段と前記データ選択手段におけるデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリもしくは大容量データ格納装置に格納させる制御手段と、

前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段と、
を有することを特徴とする知識発見方式。

【請求項１３】　前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに格納することを特徴とする請求項１２に記載の知識発見方式。

【請求項１４】　前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに格納し、前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする請求項１２に記載の知識発見方式。

【請求項１５】　前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに格納することを特徴とする請求項１２に記載の知識発見方式。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、大容量ディスクなどのデータ格納装置に格納されたデータベースから有益な知識を抽出するデータマイニングに用いて好適な知識発見方式に関する。

【０００２】

【従来の技術】従来、データマイニングまたは知識発見方式は、ディスクに蓄えられた大量のデータに対してなるべく少い回数アクセスし、読み込むことにより有効な知識発見を行なうことに主眼が置かれていた。例えば、結合ルールの抽出方式がその代表的なものであり、この方式については１９９３年発行の国際会議の予稿集「プロシーディングス　オブ　エーシーエム　シグモッド

コンファレンス　オンマネージメント　オブ　データ(Proceedings of ACM SIGMOD Conference on Management of data)」２０７頁―２１６頁に記載のアグラワル(Aggrawal)らによる論文「マイニング　アソシエーション　ルールズ　ビトウィーン セッツ オブアイテムズ　イン　ラージ　データベーセス(Mining Association Rules Between Sets of Items in Large Databases)」に詳細に記載されている。

【０００３】また、ディスクに蓄えられたデータの選択的サンプリングの方式としては、集団質問学習と呼ばれる方法が知られており、１９９２年発行の国際会議の予稿集「プロシーディングス　オブ　ザ　フィフス　アニュアル　エーシーエム　ワークショップ　オン　コンピュテーショナル　ラーニング　セオリー(Proceedingsof the fifth annual ACM workshop on computational learning thoery)」２８７頁―２９４頁に記載のセング(Seung)らの論文「クエリー　バイ　コミッティ(Query by Committee)」に詳細が記載されている。この方法では、ランダム化された下位学習方式に複数回予測をさせ、その不一致度が大きい点のみを選択するということを行う。この方法に用いられる下位学習方式はランダム化された高性能な学習方式であることが前提になっている。

【０００４】一方、性能の比較的低い学習方式の精度を増強するという文脈において、与えられたデータから繰り返し再サンプリングを行ない、そのデータを用いて学習させて得られた複数の仮説を統合することにより学習精度を向上させる一連の技術が近年注目されている。こうした技術の代表的な手法はバッギング並びにブースティングであり、これらの技術は実験的に高い性能を有することが確かめられている。バッギングの手法については、１９９４年発行のカリフォルニア大学バークレー校の技術報告書４２１に記載されたブライマン(Breiman)の論文「バッギング　プレディクターズ(Bagging Predictors)」に記載されている。

【０００５】ブースティングの手法については、１９９５年発行の国際会議予稿集「プロシーディングス　オブ　ザ　セカンド　ヨーロピアン　コンファレンス　オン　コンピュテーショナル　ラーニング　セオリー(Proceedings of the second european conference on computational learning thoery)」２３頁－３７頁に記載のフロインド(Freund)とシャピレ(Shapire)の論文「ア　デシジョン　セオレティック　ジェネラライゼーション　オブ　オンライン　ラーニング　アンド　アンアプリケーション　トゥー　ブースティング(A decision-theoretic generalization of on-line learning and an application to boosting)」に記載されている。上記の集団質問学習方式は、テキスト分類等の分野において、人手で分類しラベルづけする文書を選択する問題等に適用されている。また、バッギング技術やブースティング技術は、受動学習における精度増強の目的で利用されてきている。

【０００６】

【発明が解決しようとする課題】ところで、上述したように、従来のデータマイニング手法では、効率的なデータマイニングのために、選択的サンプリング方式を用いる場合、下位学習方式はランダム化された高性能な学習方式でなければならないという問題がある。また、バッキング技術あるいはブースティング技術を用いて性能の比較的低い学習方式の精度を上げることはできるが、両方の技術の利点を合わせて用いることができる選択的サンプリング方式がないという問題がある。

【０００７】この発明は、上記の点に鑑みてなされたもので、その目的は、大容量データ格納装置に格納された大量データからサンプル候補データを読み出し、比較的精度の低い学習方式を用いて学習させ、その結果により情報量の大きいデータを選択してメインメモリに読み込むことにより、効率的に高精度のラベル予測規則を発見する汎用的な知識発見方式を提供することにある。

【０００８】

【課題を解決するための手段】上記の目的を達成するために、請求項１に記載の発明は、大容量データ格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、前記メインメモリに格納されているデータからサンプリングして作成した複数の部分データ集合を前記学習アルゴリズムへ訓練データとして入力して学習させ、複数の仮説を得る学習手段と、該学習手段により得られた複数の仮説を用いて前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を１つあるいは複数選択し、前記メインメモリに追加して格納するデータ選択手段と、前記学習手段と前記データ選択手段による選択的なデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリに格納させる制御手段と、前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段とを有することを特徴とする。

【０００９】請求項２に記載の発明は、請求項１に記載の知識発見方式において、前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに追加して格納することを特徴とする。請求項３に記載の発明は、請求項１に記載の知識発見方式において、

前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに追加して格納し、前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする。

【００１０】請求項４に記載の発明は、請求項１に記載の知識発見方式において、前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに追加して格納することを特徴とする。請求項５に記載の発明は、請求項１または２に記載の知識発見方式において、前記データ選択手段は、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、前記予測値の分散の小さい候補点を１つまたは複数選択して前記メインメモリの格納データから削除することを特徴とする。

【００１１】請求項６に記載の発明は、請求項１または３に記載の知識発見方式において、前記データ選択手段は、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、多値予測の場合、前記予測値のマージンの大きい候補点を１つまたは複数選択し、実数値予測の場合、前記予測値の重み付き分散値の小さい候補点を１つまたは複数選択し、前記メインメモリの格納データから削除することを特徴とする。請求項７に記載の発明は、前記データ選択手段は、請求項１または４に記載の知識発見方式において、前記候補点を選択し前記メインメモリに格納するとき、既に前記メインメモリに格納されているデータに対して、前記複数の仮説を用いて関数値の予測値を求め、多値予測の場合、前記予測値の誤差マージンの大きい候補点を１つまたは複数選択し、実数値予測の場合、前記予測値の重み付き予測誤差の小さい候補点を１つまたは複数選択し、前記メインメモリの格納データから削除することを特徴とする。

【００１２】請求項８に記載の発明は、大容量データ格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、前記メインメモリに格納されたデータを訓練データとして入力して学習し、学習した仮説を大容量データ格納装置もしくはメインメモリに格納する学習手段と、該学習手段により得られた過去の複数の仮説を用いて、前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を複数選択し、前記メインメモリに格納するデータ選択手段と、前記学習手段と前記データ選択手段におけるデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリもしくは大容量データ格納装置に格納させる制御手段と、前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段と、を有することを特徴とする。

【００１３】請求項９に記載の発明は、請求項８に記載の知識発見方式において、前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに格納することを特徴とする。請求項１０に記載の発明は、請求項８に記載の知識発見方式において、前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに格納し、前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする。

【００１４】請求項１１に記載の発明は、請求項８に記載の知識発見方式において、前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに格納することを特徴とする。

【００１５】請求項１２に記載の発明は、大容量データ

格納装置に格納されたデータベースからサンプリングしたデータを計算機のメインメモリに読み込み高次知識を抽出するデータマイニングの知識発見方式において、学習アルゴリズムおよび該学習アルゴリズムに学習させるデータを入力する入力手段と、前記メインメモリに格納されているデータからサンプリングして作成した複数の部分データ集合を前記学習アルゴリズムへ訓練データとして入力して学習させ、複数の仮説を得る学習手段と、該学習手段により得られた過去の複数の仮説を用いて、前記大容量データ格納装置から読み出された複数のサンプル候補点に対する関数値の予測を行い、求めた予測値に基づき候補点の情報量を推定し、情報量の大きい候補点を複数選択し、前記メインメモリに格納するデータ選択手段と、前記学習手段と前記データ選択手段におけるデータ格納および知識発見の処理を予め定めた停止条件が満たされるまで繰り返し、その結果得られた複数の仮説を最終仮説として前記メインメモリもしくは大容量データ格納装置に格納させる制御手段と、前記複数の仮説間の平均または重み付き平均をもって前記入力部に入力したラベル未知のデータに対してラベル値を予測する予測手段とを有することを特徴とする。

【００１６】請求項１３に記載の発明は、請求項１２に記載の知識発見方式において、前記データ選択手段は、前記学習手段により得られた複数の仮説を用いて前記サンプル候補点に対する関数値の予測を行い、求めた予測値の分散値により前記候補点の情報量を推定し、分散値の大きい候補点を１つあるいは複数選択し、前記メインメモリに格納することを特徴とする。請求項１４に記載の発明は、請求項１２に記載の知識発見方式において、前記データ選択手段は、前記学習手段により得られた各仮説の前記訓練データに対する予測誤差の関数として各仮説の重みを計算し、データ選択の際に、多値予測の場

$$S = \langle (x1, y1), \ldots, (xm, ym) \rangle$$

ここで、mはデータ数であり、ｘiはデータ中のあるレコードの予め指定された属性の組みの属性値の組み、ｙiは同一レコードの予め指定されたラベル属性の属性値である。

【００１８】次に、上位アルゴリズムについて説明する。ここで、大容量データ格納装置から一部データを選択的にメモリに格納し、リサンプリング、および学習をするという手順を繰り返す回数（ステージ数）をN、各ステージにおけるリサンプリングの回数をT、メインメモリに格納すべきデータ点を選ぶ際の候補点の数をR、その中から実際にメインメモリに格納する点の数をQとする。上位アルゴリズムは、以下の手順１から手順３をN回繰り返す。もしくは、手順１の代わりに手順１’、また、手順３の代わりに手順３’を使用する。

【００１９】（手順１）メインメモリに格納されたデータからリサンプリングにより得られた複数のデータ集合S1,..., STに対して、下位学習アルゴリズムを走らせて合、各候補点に対して予測値を求める前記仮説の重みの総和を算出し、最も大きい重みの総和と次に大きい重みの総和の差分であるマージンを求め、このマージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記仮説による予測値の重み付き分散値の大きい候補点を１つまたは複数選択して前記メインメモリに格納し、前記予測手段は、前記仮説の重みによる重み付き平均をもって最終仮説による予測を行うことを特徴とする。請求項１５に記載の発明は、請求項１２に記載の知識発見方式において、前記データ選択手段は、前記大容量データ格納装置から読み出されたデータに含まれるデータ候補点の真のラベルを利用して、多値予測の場合、前記仮説の予測誤差を用いて誤差マージンを算出し、該誤差マージンの小さい候補点を１つまたは複数選択し、あるいは、実数値予測の場合、前記予測値の重み付き予測誤差の大きい候補点を１つないし複数選択して前記メインメモリに格納することを特徴とする。

【００１７】

【発明の実施の形態】以下、先ず、この発明の基本的考え方を説明する。この発明のアルゴリズムは、入力として与えられる下位学習アルゴリズムと、これを用いて、選択的サンプリングを行いながら知識発見を行なう上位アルゴリズムからなる。下位学習アルゴリズムの機能は、入力されたデータＳから学習を行い仮説を出力することと、データＳの１つのデータ点（属性値の組み）ｘに対して、学習により得られた仮説を使用して、そのラベルｙの予測値を出力することである。ここで使用する学習アルゴリズムは、高度な学習性能を持つ学習アルゴリズムは必要なく、例えば、決定木を学習するアルゴリズムや階層型のニューラルネットワークの学習アルゴリズムなどを用いることができる。ラベル付きの学習データＳは次式で表される。

（１）

仮説H1,..., HTを得る。ここで、データ $S = \langle (x1, y1), \ldots, (xm, ym) \rangle$ からのリサンプリングとは、例えばデータＳから一様分布によりサンプリングする（即ち、データＳ中の各データを１／ｍの確率で取り出す）ことを一定回数繰り返すことを言う。また、仮説とは、任意の属性値の組みに対して、そのラベルの予測値を規定するような表現形である。

（手順１’）メインメモリに格納されたデータＳに対して、下位アルゴリズムＡを走らせて仮説Ｈを得る。

【００２０】（手順２）Ｒ個の候補サンプル点ｘ1,..., ｘRを大容量データ格納装置からメインメモリに読み込み、その中から情報量の大きいＱ個のサンプル点を選択しメインメモリに格納する。他の点についてはメインメモリから削除する。情報量の大きいサンプル点の選択においては、各候補点に対し、上記手順１で得られたＴ個の仮説によりそのラベルの予測をさせて、その予測値の分散が大きいＱ個の点を選択する。

11

（手順３）過去のデータに、手順２において得られたデータを加える。

（手順３'）過去のデータを、ステップ２において得られたデータに置き換える。

【００２１】なお、上記のリサンプリングの手順を逐次的に変化する分布を用いて行なうことも可能である。例えば、精度増強法としてブースティング方式を用いる場合には、直前回の部分サンプルに対して得られた仮説の予測値が誤るようなデータの分布を逐次的に変化させながらリサンプリングを行なう。この場合、各仮説の入力訓練データに対する予測誤差の関数として各仮説の重みを計算し、それらの重みによる重み付き平均をもって最

12

終仮説による予測を行なう。そして、データ選択の際の情報量の推定にもこれらの重みを用いる。例えば、発見すべき知識表現が多値関数の場合には、各候補データに対して、各予測値を予測する仮説の重みの総和を算出し、もっとも大きい重みの総和と次に大きい重みの総和との差分（以下、この量をマージンと呼ぶ）を用いて、情報量を測ることができる。

【００２２】上記の手順をより詳細に説明する。データ点ｘに対するマージンＭ（ｘ）は以下のように定義される。

【数１】

$$M(x)=\sum_{Hi(x)=Ymax} w_i - \max_{Y\neq Ymax} \sum_{Hi(x)=Y} w_i \qquad \cdots\cdots(2)$$

ここで、Ｈｉ（ｘ）は仮説Ｈｉのデータ点ｘに対する予測値、ｗｉは、仮説Ｈｉに付された重みを指す。また、ｙｍａｘは重み総和最大の予測値とする。即ち、

【数２】

$$y_{max}=\arg\max_{Y} \sum_{Hi(x)=Y} w_i \qquad (3)$$

となる。この場合、マージン最小の点が情報量最大の点と推定される。なお、各仮説の重みを導入しない場合においても、マージンの概念は各仮説の重みを１とすることによって拡張可能なので、上記の手順によってマージン最小の点を情報量最大の点として選択することが可能である。

【００２３】発見すべき知識表現が実数関数の場合には、各候補データに対して、予測値の分散値の重み付き平均値を用いて、情報量を測ることができる。分散値の重み付き平均Ｖ（ｘ）は、次式で表される。

【数３】

$$V(x)=\sum_{i=1,\ldots,T} w_i |H_i(x)-\overline{H}_i(x)| \qquad (4)$$

ここで、Ｈｉ（ｘ）にバーを付したものはｘのラベル予測値の平均値である。分散値の重み付き平均値の大きい方が候補データの情報量は大きいとみなす。

【００２４】以上、各仮説の予測値の分散値やマージン値によって情報量を算出することを説明したが、各データ点に対する正しいラベル値を用いて算出可能な、仮説の予測誤差を利用することも可能である。例えば、発見すべき知識表現が多値関数の場合、（２）式のマージンの定義を修正して、次式で表される誤差マージンＭ'（ｘ）を定義して誤差マージン最小のデータ点を情報量最大とすることができる。

【数４】

$$M'(x)=\sum_{Hi(x)=Y^*} w_i - \max_{y\neq y^*} \sum_{Hi(x)=Y} w_i \qquad (5)$$

ここで、ｙ＊はデータｘの真のラベル値を表す。

【００２５】また、発見すべき知識表現が実数関数の場合には、各仮説のデータ点ｘに対する予測誤差の重み付き平均値Ｖ'（ｘ）として算出することが可能である。

$$V'(x)=\sum_{i=1,\ldots,T} w_i |H_i(x)-y^*| \qquad (6)$$

例えば、誤差として二乗誤差を用いればＶ'（ｘ）は次式から算出される。

【数５】

以上の定義によれば、Ｍ'（ｘ）はＭ（ｘ）の定義中の予測値モードを真のラベル値で置き換えることにより得られ、Ｖ'（ｘ）はＶ（ｘ）の定義中の予測平均値を真の値で置き換えることにより得られる。一般には、二乗誤差以外の誤差の測度、例えば絶対誤差を用いることも可能である。

【００２６】さらに、上記の上位アルゴリズムにおいて、メインメモリ量に限度があり、繰り返し選択したデータの蓄積がその限度を超過してしまう場合や、繰り返し選択し、メインメモリに格納したデータの一部が学習に不要となる場合、あるいはそれらデータを学習するには計算時間がかかり全てを学習することが実時間で不可

能な場合に、上記誤差マージンや　Ｖ′（ｘ）を利用して、メインメモリのデータを削除することが可能である。これは、メインメモリに格納されたデータのうち、情報量の小さなデータを削除することにより達成される。例えば、発見すべき知識表現が多値関数の場合には、上記誤差マージンを利用して、誤差マージンが比較的大きな点を１つまたは複数、メインメモリから削除する。また、発見すべき知識表現が実数関数の場合には、上記Ｖ′（ｘ）を算出し、Ｖ′（ｘ）の小さなデータから削除する。

【００２７】以下、図面を参照してこの発明の実施の形態について説明する。図１は、同実施形態による知識発見方式のプログラム１００の構成を示す図である。プログラム１００は、下位学習アルゴリズムと訓練データとして学習させるデータを入力する入力部１０７と、精度増強部１０３と、入力部１０７へ入力されたラベル未知のデータに対しラベル値を予測する予測部１０２と、予測部１０２が予測したラベル値を出力する出力部１０１と、精度増強部１０３における繰り返し処理を制御する制御部１０６とから構成される。

【００２８】図１の各部について説明する。精度増強部１０３は、高速メインメモリ１２０に格納されているデータから再サンプリングによって作成された複数の部分サンプルを下位学習アルゴリズムに学習させ、複数の仮説を得る学習手段１０４と、学習手段１０４により得られた複数の仮説を用いて、大容量データ格納装置より読み出された候補データ各々の情報量の推定を行ない、情報量の大きいデータのみを高速メインメモリ１２０に追加格納するデータ選択手段１０５からなる。大容量データ格納装置１３０、高速メインメモリ１２０およびＣＰＵ（中央演算処理装置）１１０は、プログラム１００を実行する計算機を構成する要素である。また、精度増強部１０３のその他の動作例を説明する。精度増強部１０３は、メインメモリ１２０に格納されているデータから学習を行い、学習した仮説をメインメモリ１２０もしくは大容量データ格納装置１３０に格納する学習手段１０４と、大容量データ格納装置１３０より読み出された候補データに対し、過去に学習された仮説を用いて情報量の推定を行い、情報量の大きなデータのみをメインメモリ１２０に格納するデータ選択手段１０５からなる。

【００２９】次に、上記構成による知識発見方式の動作を図面を参照して説明する。図２は、本発明の実施例の動作の流れを示す図である。先ず、下位学習アルゴリズム、例えば、決定木を学習するアルゴリズムなどを入力部１０７へ入力する（ステップＳ２０１）。次に、ステップＳ２０２で、学習手段１０４は、この時点で高速メインメモリ１２０に格納されているデータからリサンプリングによって部分データ集合を作成する。この部分データ集合を入力部１０７を介して入力し、前記下位学習アルゴリズムにより学習させ、仮説を得る（ステップＳ２０３）。

【００３０】次に、リサンプリング回数ｉと停止条件となるリサンプリング回数Ｔを比較し、リサンプリング回数ｉが所定の回数Ｔを越えないとき（ステップＳ２０４；ＮＯ）、ステップＳ２０２に戻り、上記の処理を繰り返す。リサンプリング回数ｉが所定の回数Ｔを越えると（ステップＳ２０４；ＹＥＳ）、データ選択手段１０５は、最終的に得られた前記仮説を用いて、大容量データ格納装置１３０から読み出された候補データの情報量の推定を行い情報量の大きいデータを選択する（ステップＳ２０５）。データ選択手段１０５により選択されたデータを高速メインメモリ１２０に格納し、既に格納してあるデータに加える（ステップＳ２０６）。

【００３１】次に、ステップＳ２０７において、ステージ数ｊを停止条件となるステージ数Ｎと比較し、ステージ数ｊが所定の回数Ｎを越えないとき（ステップＳ２０７；ＮＯ）、ステップＳ２０２から処理を繰り返す。ステージ数ｊが所定の回数Ｎを越えたとき（ステップＳ２０７；ＹＥＳ）、知識発見の過程を終了し、得られた規則を最終仮説として出力する。図２と同様に、図３も本発明の実施例の動作の流れを示す図である。図３においては、まず下位学習アルゴリズムを入力部１０７へ入力する(ステップＳ３０１)。次に、現在メインメモリに格納されている学習データを入力して前記下位学習アルゴリズムに学習させ、学習した仮説をメインメモリ１２０もしくは大容量データ格納装置１３０に格納する(ステップＳ３０２)。次に、データ選択手段１０５は、これまで過去に得られた仮説を用いて、大容量データ格納装置１３０から読み出された候補データの情報量を推定し、情報量の大きなデータを選択し(ステップＳ３０３)、メインメモリに格納する(ステップＳ３０４)。最後に、ステップＳ３０５において、ステージ数ｊを停止条件となるステージ数Ｎと比較し、ステージ数ｊが所定の回数Ｎを越えない時(ステップＳ３０５：ＮＯ)、ステップＳ３０２から処理を繰り返す。ステージ数ｊが所定の回数Ｎを越えた時(ステップS.３０５：ＹＥＳ)、知識発見の過程を終了し、得られた規則を最終仮説として出力する。

【００３２】

【発明の効果】以上説明したように、本発明の知識発見方式によれば、ディスクなどの大容量データ格納装置に格納された膨大な量のデータから情報量の大きい部分だけを選択的にサンプリングして得られた比較的少数のデータから効率的に、高精度で未知データのラベル予測を行なう規則を発見することが可能になり、データマイニングの実効性を高めるという効果が得られる。また、データの選択的サンプリングにより少数のデータをメインメモリに読み込み知識発見を行うことからメインメモリに制限のある計算機でもデータマイニングを行うことができるという効果が得られる。

15

【図面の簡単な説明】

【図１】　この発明の一実施の形態の構成を示す図である。

【図２】　同実施の形態の動作の流れを示す図である。

【図３】　同実施の形態の動作の流れを示す図である。

【符号の説明】

１００　知識発見プログラム

16

１０１　出力部

１０２　予測部

１０３　精度増強部

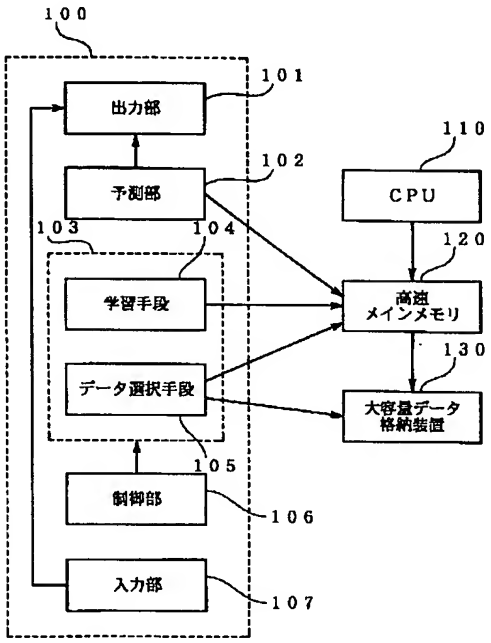
１０４　学習手段

１０５　データ選択手段

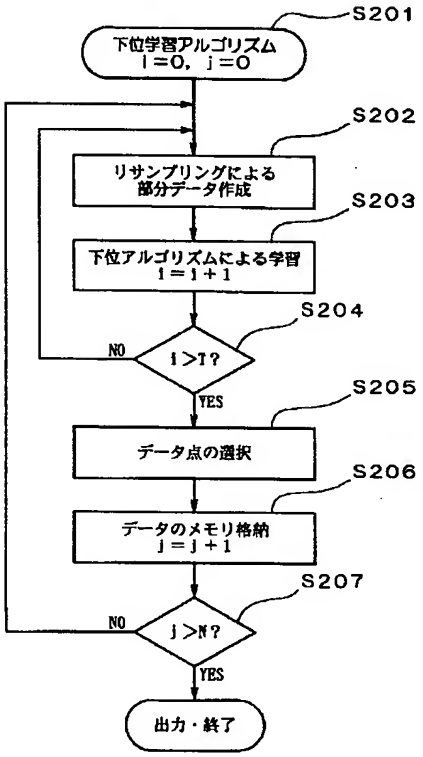
１０６　制御部

１０７　入力部

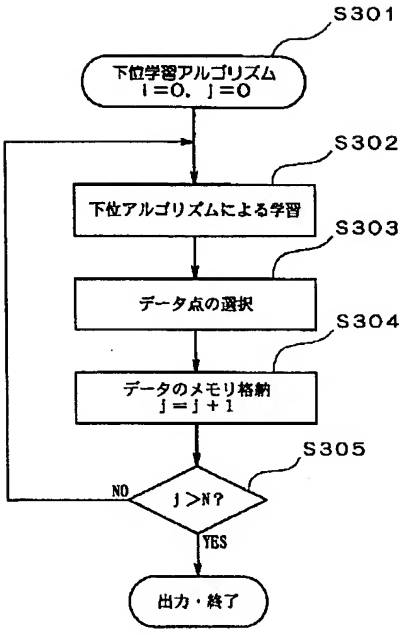
【図１】



【図２】



【図３】


S301
下位学習アルゴリズム
i＝0，j＝0
S302
下位アルゴリズムによる学習
S303
データ点の選択
S304
データのメモリ格納
j＝j＋1
S305
NO
j＞N？
YES
出力・終了